On asymptotic efficiency of multivariate version of Spearman's rho

Alexander Nazarov* and Natalia Stepanova†

Abstract

A multivariate version of Spearman's rho for testing independence is considered. Its asymptotic efficiency is calculated under a general distribution model specified by the dependence function. The efficiency comparison study that involves other multivariate Spearman-type test statistics is made. Conditions for Pitman optimality of the test are established. Examples that illustrate the quality of the multivariate Spearman's test are included.

Key words: Spearman's rho, multivariate rank statistic, test of independence, Pitman efficiency, U-statistic, Lagrange principle

AMS 2000 subject classifications: primary 62G10, 62G20

1 Introduction

Testing for independence among the components of m-variate vector is an important statistical problem. There is an extensive statistical literature on this topic. Over the last two decades a variety of new multivariate measures of association have been suggested, including those based on ranks, and their properties have been studied.

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})$, $m \geq 2$, $i = 1, \ldots, n$, be independent random vectors with absolutely continuous cdf F and marginal cdfs $F_1, \ldots, F_m$. Denote by R_{ij} the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. In case of bivariate random sample, when $m = 2$, a commonly used statistic for testing the hypothesis of independence, $H_0 : F \equiv F_1 F_2$, is Spearman's correlation coefficient [21]

$$\rho_n = \frac{12}{n^2 - 1} \left\{ n^{-1} \sum_{i=1}^{n} R_{i1} R_{i2} - \left(\frac{n+1}{2} \right)^2 \right\}$$

*Department of Mathematics and Mechanics, St.Petersburg State University, 28 Universitetskii Prospekt, St. Petersburg, 198504, Russia

†School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada

that estimates the functional

$$\rho(F) = 12 \int F dF_1 dF_2 - 3. \tag{1}$$

Among various multivariate extensions of Spearman's rho available in the statistical literature, the following three statistics seem to be quite popular (see, for example, [9], [18] [20], [22]):

$$S_{m,n} = \frac{1}{C_m} \left\{ n^{-1} \sum_{i=1}^{n} \prod_{k=1}^{m} (n+1-R_{ij}) - \left(\frac{n+1}{2}\right)^m \right\} \tag{2}$$

$$W_{m,n} = \frac{1}{C_m} \left\{ n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{m} R_{ij} - \left(\frac{n+1}{2}\right)^m \right\}, \tag{3}$$

$$V_{m,n} = \frac{12}{n^2-1} \left\{ \binom{m}{2}^{-1} \sum_{1 \le j < j' \le m} n^{-1} \sum_{i=1}^{n} R_{ij} R_{ij'} - \left(\frac{n+1}{2}\right)^2 \right\}. \tag{4}$$

where $C_m = n^{-1} \sum_{i=1}^{n} i^m - ((n+1)/2)^m$ is a normalizing factor.

Statistic (4) is simply the average pair-wise Spearman's rho [11, Ch. 6] that estimates [9]

$$\nu_m(F) = 12 \left\{ \binom{m}{2}^{-1} \int \sum_{j<j'} F_j F_{j'} dF \right\} - 3,$$

Statistics (2) and (3) are natural generalization of Spearman's rho, as they are sample counterparts of the functionals

$$s_m(F) = \frac{1}{d_m} \left\{ \int F dF_1 \dots dF_m - c_m \right\},$$
$$w_m(F) = \frac{1}{d_m} \left\{ \int F_1 \dots F_m dF - c_m \right\},$$

where $c_m = 2^{-m}$, $d_m = (m+1)^{-1} - 2^{-m}$, respectively. The correspondence between $S_{m,n}$, the main object under investigation in this paper, and $s_m(F)$ is easy to see. Indeed, let F_n be the multivariate empirical cdf that corresponds to F, and let $F_{j,n}$ be the marginal empirical cdfs based on $X_{1j}, \dots, X_{nj}$, $j = 1, \dots, m$. Then

$$R_{ij} = \sum_{k=1}^{n} \mathbb{I}(X_{kj} \le X_{ij}) = n F_{j,n}(X_{ij}) = (n+1) F^*_{j,n}(X_{ij}),$$

where $F^*_{j,n} = (n/(n+1)) F_{j,n}$ are the modified empirical cdfs. Therefore $S_{m,n}$ can be written in the form

$$S_{m,n} = \frac{(n+1)^m}{C_m} \left(\int \prod_{j=1}^{m} (1 - F^*_{j,n}) dF_n - \frac{1}{2^m} \right),$$

where, taking into account that $\sum_{i=1}^{n} i^m \sim n^{m+1}/(m+1)$, as $n \to \infty$, we have

$$(n+1)^m / C_m \sim (1/(m+1) - 1/2^m)^{-1} = 1/d_m.$$

Thanks to the Glivenko–Cantelli theorem (see, for example, [3, Sec. I.4, Th. 1]) the closeness of $S_{m,n}$ and $s_m(F)$ is now immediately seen by noting that

$$\int_{\mathbb{R}^m} \prod_{j=1}^{m} (1 - F_j(x_j)) dF(x_1, \ldots, x_m) = \int_{\mathbb{R}^m} F(x_1, \ldots, x_m) \prod_{j=1}^{m} dF_j(x_j). \tag{5}$$

Equality (5) is easy to verify by integrating by parts on the left-hand side and using the properties of a multivariate cdf.

All three measures of multivariate concordance, $\nu_m(F)$, $s_m(F)$, and $w_m(F)$, increase with respect to the *multivariate concordance ordering* introduced by Joe [9, Sec. 2], [10, Ch. 2]. This ordering is based on the concept of *positive orthant dependence* [10, Sec. 2.1]. It results from a comparison of a multivariate random vector with a random vector of independent random variables having the same univariate marginal distributions. More precisely, let F and G be two m-variate cdfs with corresponding survival functions $\bar{F}$ and $\bar{G}$, i.e. $\bar{F}(x_2, \ldots, x_m) = \mathbf{P}_F(X_1 > x_1, \ldots, X_m > x_m)$ and $\bar{G}(x_2, \ldots, x_m) = \mathbf{P}_G(Y_1 > x_1, \ldots, Y_m > x_m)$. Then G is said to be *more concordant* than F (written $F \prec_c G$) if

$$F(\mathbf{x}) \le G(\mathbf{x}) \quad \text{and} \quad \bar{F}(\mathbf{x}) \le \bar{G}(\mathbf{x}), \quad \text{for all } \mathbf{x} = (x_1 \ldots, x_m) \in \mathbb{R}^m.$$

That is, if $\mathbf{X} = (X_1, \ldots, X_m) \sim F$, $\mathbf{Y} = (Y_1, \ldots, Y_m) \sim G$, and $F \prec_c G$, then the components of $\mathbf{Y}$ are more likely than those of $\mathbf{X}$ to take on small and large values simultaneously. As shown in [9], $F \prec_c G$ implies $\nu_m(F) \le \nu_m(G)$ and $w_m(F) \le w_m(G)$. The fact that $s_m(F)$ is also increasing with respect to $\prec_c$ follows immediately from Lemma 3.3.1 of [9], the Remark below this lemma, and equality (5).

Unlike the classical problem of testing independence when $m = 2$, there is still no clear clear concept of *negative* multivariate concordance. Some "characterizations" of the negative multivariate concordance can be found, for example, in [9].

In connection with testing independence among the components of a m-variate random vector, statistics (2)–(4) were studied by several authors. One of the earliest comprehensive study related to multivariate rank statistics for testing independence can be found in [19]. The Pitman efficiency properties of the tests based on (2)–(4) are investigated in [9], [18], [22], among others. The asymptotic normality of statistics (2)–(4) is established in [20] under rather weak assumptions on the underlying distribution. The asymptotic efficiency study of the Spearman-type tests, including those based on $S_{m,n}$ and $V_{m,n}$, is conducted in [18] under various distribution models. The thorough study of Pitman efficiency properties of $W_{m,n}$ and $V_{m,n}$ is done in [22].

An interesting problem related to finding the Pitman efficiency of a test is to discover the structure of the underlying distribution for which the test is Pitman optimal. Many test statistics were suggested by their authors empirically for solving particular problems of testing hypotheses, and were supposed to work in one or another particular situation. Problems of finding the most

favourable alternatives have been studied in [2], [6], [16, Ch. 6], [17], [22], etc. In this paper, assuming one-parameter model

$$F_\theta(\mathbf{x}) = \prod_{j=1}^{m} F_j(x_j) + \theta\Omega_m(F_1(x_1), \ldots, F_m(x_m)), \quad \mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m, \tag{6}$$

where θ is a parameter of association close to zero and Ω_m is the *dependence function* defined on the unit m-cube and satisfying certain boundary and smoothness conditions, we find the most favourable alternative to independence for which the test based on $S_{m,n}$ is Pitman optimal.

In order to determine the "optimal" distribution function one has to solve a variational problem of minimization of an appropriate functional on a set of special type, depending on the structure of the test statistic. Typically, optimality conditions for tests are found by using the Lagrange multiplier rule applied to a functional on a Banach space. Under the validity of model (6), the optimality problems for the Spearman-type test statistics $W_{m,n}$ and $V_{m,n}$ have been solved in [22]. Compared to these two cases, the extreme problem related to $S_{m,n}$ is much more complicated and is reduced to solving the system of partial differential equations with non-standard boundary conditions. In Section 4.2 we provide solution to a general m-dimensional extremal problem that gives Pitman optimality conditions for the sequence $\{S_{m,n}\}_{n\geq 1}$.

In Section 2 we introduce statistical model and describe its properties. Some basic properties of the test statistic $S_{m,n}$, including its asymptotic normality in terms of the dependence function, are given in Section 3. Asymptotic efficiency study is performed in Section 4. The key result of the paper, the Theorem of Section 4.3, provides the most favourable alternative to independence for the test statistic at hand.

2 Multivariate model

2.1 Definition of the model

Suppose we observe an m-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of size n from distribution $\mathbf{P}_\theta$ on the measurable space $(\mathbb{R}^m, \mathcal{B}^m)$ indexed by a parameter $\theta \geq 0$. Then the full observation is a single observation from the product $\mathbf{P}_\theta^n$ of n copies of $\mathbf{P}_\theta$. Let F_θ be the distribution function that corresponds to $\mathbf{P}_\theta$. When testing independence among the components of a continuously distributed random vector, without loss of generality, the marginal cdfs F_j, $j = 1, \ldots, m$, can be taken uniformly distributed on the interval $[0, 1]$. Then, the statistical model is described in terms of distribution functions as the collection of probability measures $\{\mathbf{P}_\theta^n : \theta \geq 0\}$ on the sample space $(\mathbb{R}^{m\times n}, \mathcal{B}^{m\times n})$ such that

$$F_\theta(\mathbf{x}) = \prod_{j=1}^{m} x_j + \theta\Omega_m(\mathbf{x}), \quad \mathbf{x} = (x_1, \ldots, x_m) \in [0, 1]^m = I^m,\ m \geq 2, \tag{7}$$

is satisfied for sufficiently small value of θ subject to some restrictions on Ω_m. To be precise, let $\mathcal{F}_m = \{F_\theta\}$ be the class of absolutely continuous cdfs of type (7) for which, cf. [22, Sec. 2],

(**C1**) $\Omega_m(\mathbf{x}) \geq 0, \quad \mathbf{x} \in I^m$,

(**C2**) $\Omega_m(\mathbf{x})|_{x_k=0} = 0, \quad \Omega_m(1,\ldots,1,x_k,1,\ldots,1) = 0, \quad x_k \in [0,1], \quad 1 \leq k \leq m$,

(**C3**) $\Omega_m(\mathbf{x})|_{x_k=1} = \Omega_{m-1}(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_m), \quad 1 \leq k \leq m$,

(**C4**) there exists a non-zero mixed derivative

$$\omega_m(\mathbf{x}) = \frac{\partial^m \Omega_m(\mathbf{x})}{\partial x_{i_1} \ldots \partial x_{i_m}}, \quad \text{for } \lambda_m\text{-almost all } \mathbf{x} \in I^m,$$

such that $\omega_m \in \mathbf{L}_2(I^m)$, where l_m is the Lebesgue measure on $(\mathbb{R}^m, \mathcal{B}^m)$ and $(i_1,\ldots,i_m)$ is an arbitrary permutation of the set $\{1,\ldots,m\}$.

Due to (**C1**), boundary conditions (**C2**), and the consistency property (**C3**), for sufficiently small θ, all the properties of a multivariate cdf are satisfied. The regularity condition **(C4)** implies local asymptotic normality of the sequence of models $\{\mathbf{P}_\theta^n : \theta \geq 0\}$ at $\theta = 0$ (see Section 2.2 for details). In the sequel, m-variate sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ is assumed taken from distribution for which the cdf $F_\theta(\mathbf{x})$ belongs to $\mathcal{F}_m$, $m \geq 2$. The symbols $\mathbf{E}_\theta$ and $\mathbf{Var}_\theta$ (with index n omitted) are used below to denote the expectation and the variance with respect to $\mathbf{P}_\theta^n$.

We are interested in testing the hypothesis of independence

$$H_0 : \ \theta = 0$$

against the one-sided alternative

$$H_1 : \theta > 0.$$

In the case $m = 2$, model (7) was first studied by Farlie [4] and appeared later in a number of publications (see [22, Sec. 1] for references), sometimes with a specific choice of dependence function. Considered under assumptions (**C1**)–(**C4**), model (7) is an extension of the Farlie model to the multivariate case.

2.2 Local asymptotic normality of the model

Recall that a sequence of statistical models is *locally asymptotically normal* (LAN) if it converges to a Gaussian model whose properties are well known [8, Sec. 2], [23, Sec. 7].

Let f_θ be the density of $\mathbf{P}_\theta$ with respect to λ_m, that is,

$$f_\theta(\mathbf{x}) = 1 + \theta\omega_m(\mathbf{x}), \quad \mathbf{x} \in I^m, \quad m \geq 2,$$

and denote by $\dot{f}_\theta(\mathbf{x})$ its partial derivative with respect to θ. The true statistical difficulty is to distinguish between the null hypothesis and the alternative when θ is small, typically "of size

$O(n^{-1/2})$." Therefore we introduce a *local parameter* $h = \sqrt{n}\,\theta$, and consider a local statistical experiment indexed by h:

$$(\mathbf{X}_1, \ldots, \mathbf{X}_n) \sim \{\mathbf{P}^n_{h/\sqrt{n}} : h \geq 0\}.$$

Our attention will be focused on the performance of the test based on $S_{m,n}$ at alternatives

$$H_{1n} : h > 0$$

converging, as $n \to \infty$, to the null hypothesis

$$H_0 : h = 0.$$

Let $\Delta_{n,\theta}$ be a random vector such that $\Delta_{n,\theta} \xrightarrow{\mathrm{d}} \mathcal{N}(0, I_\theta)$, where for $\theta = \theta_n = h/\sqrt{n}$,

$$I_\theta = \mathbf{E}_\theta \left(\frac{\partial}{\partial \theta} \log(d\mathbf{P}_\theta / d\lambda_m) \right)^2 = \int_{I^m} \frac{\dot{f}^2_\theta(\mathbf{x})}{f_\theta(\mathbf{x})}\, d\mathbf{x}, \quad \theta \geq 0,$$

is the Fisher information in the parametric family $\{f_\theta(\mathbf{x}),\, \theta \geq 0\}$. Thanks to Theorem 1.1 of [8], under the regularity condition (**C4**), the sequence of statistical experiments $\{\mathbf{P}^n_{h/\sqrt{n}} : h \geq 0\}$ is LAN at the point $h = 0$, that is, for any $h \geq 0$

$$\log \frac{d\mathbf{P}^n_{h/\sqrt{n}}}{d\mathbf{P}^n_0} = h\Delta_{n,0} - \frac{1}{2}h^2 I_0 + o_{\mathbf{P}^n_0}(1), \quad n \to \infty. \tag{8}$$

Under local asymptotic normality

$$\log \frac{d\mathbf{P}^n_{h/\sqrt{n}}}{d\mathbf{P}^n_0} \xrightarrow{\mathrm{d}} \mathcal{N}\left(-\frac{1}{2}h^2 I_0, h^2 I_0 \right), \quad n \to \infty,$$

and hence the sequences of distributions $\{\mathbf{P}^n_{h/\sqrt{n}}\}$ and $\{\mathbf{P}^n_0\}$ are mutually contiguous (see [23, Sec. 7.5]). This fact allows us to obtain, by means of Le Cam's third lemma [23, Sec. 6.7], limit distribution of $S_{m,n}$ under the sequence of alternatives H_{1n}, once the limit distribution under H_0 is known. Another useful consequence of the local asymptotic normality is the existence of an upper bound on the asymptotic power function of the test. This makes it possible to establish the conditions for asymptotic optimality of the test statistic $S_{m,n}$ (see [23, Ch.15]).

3 Basic properties and asymptotic normality

In this section we list some basic properties of the test statistic $S_{m,n}$. First, note that $S_{m,n}$ is symmetric in m variables. It is normalized so that its value is 1 when $R_{i1} = R_{i2} = \ldots = R_{im}$, or equivalently, $F_\theta = \min(F_1, \ldots, F_m)$ (perfect positive dependence), and its expected value under H_0 is zero. The lower bound of $s_m(F)$ is equal to

$$s_m\left(\max(F_1 + \ldots + F_m - m + 1, 0)\right) = \frac{2^m(m+1)}{2^m - (m+1)} \left\{ \frac{1}{(m+1)!} - \frac{1}{2^m} \right\},$$

which is -1 for $m = 2$ and is greater than -1 for $m \geq 3$. The lower bound is an increasing function of m tending to zero as m gets larger. Hence $S_{m,n}$, the sample version of $s_m(F)$, also exceeds -1 and its lower bound tends to zero as m increases. For this reason, it is appropriate to use the statistic $S_{m,n}$ for testing the hypothesis independence $H_0 : \theta = 0$ against the one-sided alternative $H_1 : \theta > 0$ only. The non-symmetry between the upper and lower bounds is due to the "curse if dimensionality" and is partly explained by the inequality [10, Lemma 3.8]

$$\max(F_1 + \ldots + F_m - m + 1, 0) \leq F_\theta \leq \min(F_1, \ldots, F_m),$$

where in contrast to the Fréchet upper bound, $\min(F_1, \ldots, F_m)$, the Fréchet lower bound, $\max(F_1 + \ldots + F_m - m + 1, 0)$ is generally not a cdf, except for the case $m = 2$. Through the curse of dimensionality, the concepts of perfect positive and perfect negative dependence lose the symmetry of the two-dimensional case.

There exist a variety of theorems on asymptotic normality of multivariate linear rank statistics. A unifying approach to these various results is given, for example, in [19]. In particular, Theorem 2 of [19] implies that $S_{m,n}$ is asymptotically normally distributed. For our purpose, however, it is more convenient to establish asymptotic normality of $S_{m,n}$ through the correspondence between $S_{m,n}$ and a closely related U-statistic.

The multivariate rank statistic $S_{m,n}$ is asymptotically equivalent to a $(m+1)$-dimensional U-statistic, $U_{m,n}$, based on $\int F_\theta dF_1 \ldots dF_m$. The kernel of the U-statistic comes from symmetrizing $\mathbb{I}(X_{m+1,j} < X_{jj}, j = 1, \ldots, m)$, cf. [22, eq. (3.4)]:

$$U_{m,n} = \binom{n}{m+1}^{-1} \sum_{1 \leq i_1 < \ldots < i_{m+1} \leq n} g(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m+1}}),$$

where

$$g(\mathbf{X}_1, \ldots, \mathbf{X}_{m+1}) = \frac{1}{(m+1)!} \sum_{(i_1, \ldots, i_{m+1})} (\mathbb{I}(X_{i_{m+1},1} < X_{i_1,1}, \ldots, X_{i_m,m} < X_{i_m,m}) - c_m)/d_m,$$

and the summation is extended over all permutations $(i_1, \ldots, i_{m+1})$ of $\{1, \ldots, m+1\}$.

The following result establishes "locally uniform" asymptotic normality of $U_{m,n}$. It will be used for calculating the *slope* (or *efficacy*) of the test statistic $S_{m,n}$ whose limit distribution coincides with that of $U_{m,n}$.

Lemma 1. *If $F \in \mathcal{F}_m = \{F_{h/\sqrt{n}}\}$, then* for all $h \geq 0$,

$$\frac{\sqrt{n}(U_{m,n} - \mu_m(h))}{\sigma_m(h)} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \quad n \to \infty,$$

where

$$\begin{aligned} \mu_m(h) &= \frac{2^m(m+1)}{2^m - (m+1)} h \int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x}, \\ \sigma_m^2(h) &= \sigma_m^2(0) = \frac{(m+1)^2}{(2^m - (m+1))^2} \left(\left(\frac{4}{3}\right)^m - \frac{m}{3} - 1 \right). \end{aligned}$$

Proof. For $\theta \geq 0$, put

$$\eta_m(\theta) = \mathbf{E}_\theta \Psi_\theta^2(X_1) - \left(\mathbf{E}_\theta U_{m,n}\right)^2, \quad \Psi_\theta(\mathbf{x}) = \mathbf{E}_\theta g(\mathbf{X}_1, \ldots, \mathbf{X}_{m+1}) | \mathbf{X}_1 = \mathbf{x}).$$

By the CLT for U-statistics (see, for example, [13, Sec. 4.2]) for all $\theta \geq 0$

$$n^{1/2}((m+1)^2 \eta_m(\theta))^{-1/2}(U_{m,n} - \mathbf{E}_\theta U_{m,n}) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \quad n \to \infty, \tag{9}$$

provided $\eta_m(\theta) > 0$ and $\mathbf{E}_\theta \Psi_\theta^2(\mathbf{X}_1, \ldots, \mathbf{X}_{m+1}) < \infty$. Note that

$$\begin{aligned}
\mathbf{E}_\theta U_{m,n} &= \mathbf{E}_\theta g(\mathbf{X}_1, \ldots, \mathbf{X}_{m+1}) \\
&= \mathbf{E}_\theta \left(\mathbb{I}(X_{m+1,1} < X_{11}, \ldots, X_{m+1,m} < X_{m,m}) - c_m \right) / d_m \\
&= \left(\int_{\mathbb{R}^m} F_\theta(\mathbf{x}) dF_1(x_1) \ldots dF_m(x_m) - c_m \right) \Big/ d_m = s_m(F_\theta).
\end{aligned}$$

In particular, $\mathbf{E}_0 U_{m,n} = 0$. Next, under H_0

$$\begin{aligned}
\Psi_0(\mathbf{x}) &= \frac{1}{(m+1)!} \sum_{(i_1, \ldots, i_{m+1})} \left\{ \mathbf{E}_0 \left(\mathbb{I}(X_{m+1,1} < X_{11}, \ldots, X_{m+1,m} < X_{m,m}) | \mathbf{X}_1 = \mathbf{x} \right) - c_m \right\} / d_m \\
&= \frac{1}{d_m} \left\{ \frac{m!}{(m+1)!} \left(\frac{1}{2^{m-1}} \sum_{j=1}^m x_j + \prod_{j=1}^m (1 - x_j) \right) - c_m \right\} \\
&= \frac{1}{2^m - (m+1)} \left\{ 2 \sum_{j=1}^m x_j + 2^m \prod_{j=1}^m (1 - x_j) - (m+1) \right\}.
\end{aligned}$$

Under model (7), the calculation of $\eta_m(0) = \mathbf{E}_0 \Psi_0^2(\mathbf{X}_1)$ can be simplified by noting that in case of independence, the vectors $\mathbf{1} - \mathbf{X}_1$ and $\mathbf{X}_1$ are equally distributed, each with i.i.d. uniform components. Therefore

$$\begin{aligned}
\eta_m(0) &= \mathbf{E}_0 \Psi_0^2(\mathbf{1} - \mathbf{X}_1) = \frac{1}{(2^m - (m+1))^2} \mathbf{E}_0 \left(2 \sum_{j=1}^m (1 - X_{1j}) + 2^m \prod_{j=1}^m X_{1j} - (m+1) \right)^2 \\
&= \frac{1}{(2^m - (m+1))^2} \left(\left(\frac{4}{3} \right)^m - \frac{m}{3} - 1 \right).
\end{aligned}$$

From this, applying (9) we get under H_0

$$\sqrt{n} \sigma_m^{-1}(0)(U_{m,n} - \mu_m(0)) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \quad n \to \infty,$$

where

$$\mu_m(0) = 0, \quad \sigma_m^2(0) = \frac{(m+1)^2}{(2^m - (m+1))^2} \left(\left(\frac{4}{3} \right)^m - \frac{m}{3} - 1 \right). \tag{10}$$

Thus, for $\theta = 0$ the lemma is proved.

Now using the "contiguity arguments" we will reduce the derivation of asymptotic normality under $\theta_n = h/\sqrt{n}$ to derivation under $\theta = 0$. First, applying the projection technique to the U-statistic $U_{m,n}$, we get

$$\sqrt{n}(U_{m,n} - \mu_m(0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_0(\mathbf{X}_i) + o_{\mathbf{P}_0}(1),$$

where

$$\psi_0(\mathbf{x}) = (m+1)\Psi_0(\mathbf{x}) = \frac{(m+1)}{2^m - (m+1)} \left\{ 2 \sum_{j=1}^{m} x_j + 2^m \prod_{j=1}^{m} (1 - x_j) - (m+1) \right\}.$$

Then, under $H_{1n} : h > 0$, Le Cam's third lemma implies (see [23, Sec. 7.5])

$$\sqrt{n}(U_{m,n} - \mu_m(0)) \xrightarrow{\mathrm{d}} \mathcal{N}\left(h\mathbf{E}_0[\psi_0(\mathbf{X}_1)\dot{l}_0(\mathbf{X}_1)], \mathbf{E}_0\psi_0^2(\mathbf{X}_1) \right),$$

where $\dot{l}_\theta(\mathbf{x}) = (\partial/\partial\theta) \log f_\theta(\mathbf{x}) = \omega_m(\mathbf{x})/(1 + \theta\omega_m(\mathbf{x}))$, $\theta \geq 0$. In other words, the statistic $U_{m,n}$ is approximately normally distributed with variance $n^{-1}\mathbf{E}_0\psi_0^2(\mathbf{X}_1) = n^{-1}\sigma_m^2(0)$, where $\sigma_m^2(0)$ is defined in (10), and mean value

$$\begin{aligned} \mu_m(h) &= h\mathbf{E}_0[\psi_0(\mathbf{X}_1)\dot{l}_0(\mathbf{X}_1)] = h \int_{I^m} \psi_0(\mathbf{x})\omega_m(\mathbf{x})d\mathbf{x} \\ &= \frac{(m+1)}{2^m - (m+1)} h \int_{I^m} \left(2 \sum_{j=1}^{m} x_j + 2^m \prod_{j=1}^{m} (1 - x_j) - (m+1) \right) \omega_m(\mathbf{x})d\mathbf{x}. \end{aligned}$$

Notice that

$$\int_{I^m} \omega_m(\mathbf{x})d\mathbf{x} = 0, \quad \int_{I^m} x_j\omega_m(\mathbf{x})d\mathbf{x} = 0, \quad 1 \leq j \leq m,$$

$$\int_{I^m} \prod_{j=1}^{m} (1 - x_j)\omega_m(\mathbf{x})d\mathbf{x} = \int_{I^m} \Omega_m(\mathbf{x})d\mathbf{x},$$

where the first two equalities are consequences of boundary conditions **(C2)**, and the third one follows from (5). Therefore

$$\mu_m(h) = \frac{2^m(m+1)}{2^m - (m+1)} h \int_{I^m} \Omega_m(\mathbf{x})d\mathbf{x}.$$

The proof is completed. □

Due to Lemma 1, the test based on $S_{m,n}$ rejects the null hypothesis of independence at level approximately α if $\sqrt{n}S_{m,n}/\sigma_m(0) > z_\alpha$, where $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the quantile of order $(1 - \alpha)$ of a standard normal distribution.

4 Asymptotic efficiency

First, we calculate the Pitman efficiency of the test statistic $S_{m,n}$. Denote by $\gamma_{m,n}(\theta)$, $\theta = h/\sqrt{n} \geq 0$, the power function of the test of level approximately α:

$$\gamma_{m,n}(\theta) = \mathbf{P}_\theta(\sqrt{n}S_{m,n}/\sigma_m(0) > z_\alpha).$$

If for a sequence of tests $\{T_n\}$ the corresponding sequence of power functions satisfies $\gamma_n(h/\sqrt{n}) \to 1 - \Phi(z_\alpha - hs)$, for every $h \geq 0$, then the sequence $\{T_n\}$ is said to have *slope* (or *efficacy*) s. A widely-recognized quantitative measure of comparison of two statistical tests is the square of the quotient of two slopes. This quantity is called the *asymptotic relative efficiency* (ARE) of the tests. Further, if the sequence of experiments $\{\mathbf{P}_\theta^n : \theta \geq 0\}$ is LAN at $\theta = 0$, then an upper bound on the slope exists [23, Th. 15.4]. This yields the relative efficiency of the test with slope s and the best test and thus allows us to determine the *absolute* quality of the former.

Lemma 1 implies that the sequence $\{S_{m,n}\}_{n\geq 1}$ is locally uniformly asymptotically normal. Then the general result on behavior of the *local limiting* power function, defined as

$$\gamma_m(h) = \lim_{n\to\infty} \gamma_{m,n}(h/\sqrt{n}), \quad h \geq 0,$$

says that γ_m depends on the sequence $\{S_{m,n}\}_{n\geq 1}$ only through the quantity $\mu'_m(0)/\sigma_m(0)$, the slope of the sequence of tests (see [23, Th. 14.7]).

4.1 Relative and absolute measures of efficiency

Next lemma gives an expression for the local limiting power function of the test at hand in terms of the dependence function Ω_m.

Lemma 2. *Assume model* (7) *and let* $\gamma_m(h) = \lim_{n\to\infty} \gamma_{m,n}\left(h/\sqrt{n}\right)$. *Then*

$$\gamma_m\left(h\right) = 1 - \Phi\left(z_\alpha - \frac{2^m}{\left((4/3)^m - m/3 - 1\right)^{1/2}} h \int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x}\right)$$

Proof. In view of Lemma 1, the proof follows immediately from Theorem 14.7 of [23]. □

From Lemma 2, the measure of efficiency for the sequence $\{S_{m,n}\}_{n\geq 1}$ is equal to

$$\left(\frac{\mu'_m(0)}{\sigma_m(0)}\right)^2 = \frac{4^m}{(4/3)^m - m/3 - 1}\left(\int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x}\right)^2. \tag{11}$$

For the multivariate Spearman-type statistics $W_{m,n}$ and $V_{m,n}$ these are (see [22, Sec. 4.1])

$$\left(\frac{\mu'_{m,W}(0)}{\sigma_{m,W}(0)}\right)^2 = \frac{4^m}{(4/3)^m - m/3 - 1}\left(\int_{I^m} \prod_{j=1}^{m} x_j \omega_m(\mathbf{x})\, d\mathbf{x}\right)^2, \tag{12}$$

and

$$\left(\frac{\mu'_{m,V}(0)}{\sigma_{m,V}(0)}\right)^2 = 144\binom{m}{2}^{-1}\left(\sum_{1\leq i<j\leq m} \int_{I_m} x_i x_j \omega_m(\mathbf{x})\, d\mathbf{x}\right)^2, \tag{13}$$

respectively. The asymptotic relative efficiency of $S_{m,n}$ relative to $W_{m,n}$ and $V_{m,n}$ is then

$$\mathrm{ARE}(S,W) \;=\; \left(\frac{\mu'_m(0)/\sigma_m(0)}{\mu'_{m,W}(0)/\sigma_{m,W}(0)}\right)^2, \quad \mathrm{ARE}(S,W) = \left(\frac{\mu'_m(0)/\sigma_m(0)}{\mu'_{m,V}(0)/\sigma_{m,V}(0)}\right)^2.$$

At this point, recall that the sequence of models $\{\mathbf{P}^n_{h/\sqrt{n}} : h \geq 0\}$ under consideration is LAN at $h = 0$. Therefore there exists an upper bound on the power function of the test (see [23, Th. 15.4]). More precisely, for all $h \geq 0$,

$$\limsup_{n\to\infty} \gamma_{m,n}(h/\sqrt{n}) \leq 1 - \Phi(z_\alpha - h\sqrt{I_0}).$$

That is, the square root of the Fisher information $I_0 = \int_{I^m} \omega_m^2(\mathbf{x})\, d\mathbf{x}$ is the largest possible slope:

$$\left(\frac{\mu_m'(0)}{\sigma_m(0)}\right)^2 \leq \int_{I^m} \omega_m^2(\mathbf{x})\, d\mathbf{x},$$

or equivalently,

$$\frac{4^m}{(4/3)^m - m/3 - 1}\left(\int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x}\right)^2 \leq \int_{I^m} \omega_m^2(\mathbf{x})\, d\mathbf{x}. \tag{14}$$

Therefore the Pitman absolute efficiency of the test based on $S_{m,n}$ is given by the formula

$$e_S(\Omega_m) = \frac{4^m}{((4/3)^m - m/3 - 1)}\left(\int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x}\right)^2 \Big/ \int_{I^m} \omega_m^2(\mathbf{x})\, d\mathbf{x}. \tag{15}$$

For a given function Ω_m, the closer the value of $e_S(\Omega_m)$ to one, the better the test based on $S_{m,n}$. Similarly, using (12) and (13)

$$e_W(\Omega_m) = \frac{4^m}{(4/3)^m - (m/3) - 1}\left(\int_{I^m} \prod_i x_i\, \omega(\mathbf{x}) d\mathbf{x}\right)^2 \Big/ \int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x}, \tag{16}$$

$$e_V(\Omega_m) = 144\binom{m}{2}^{-1}\left(\sum_{i<j} \int_{I^m} x_i x_j \omega_m(\mathbf{x}) d\mathbf{x}\right)^2 \Big/ \int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x}. \tag{17}$$

4.2 Extremal problem

We are interested in finding the most favourable alternative, determined by the dependence function $\Omega_m(\mathbf{x})$, for which the sequence of test statistics $\{S_{m,n}\}_{n\geq 1}$ has the largest possible slope. This problem is reduced to the problem of finding $\Omega_m(\mathbf{x})$ that delivers equality in inequality (14). The latter is a particular case of a general m-dimensional extremal problem treated below.

Let us introduce the space $\mathbf{C}_0^m$ of functions that are m-times continuously differentiable with respect to each variable and obey certain boundary conditions:

$$\mathbf{C}_0^m = \{\Omega \in \mathbf{C}^m(I^m) : \Omega(\mathbf{x})|_{x_j=0} = 0,\ j = 1, \ldots, m\}.$$

Define a scalar product on $\mathbf{C}_0^m$ as follows:

$$(\Omega_1, \Omega_2) = \int_{I^m} \omega_1(\mathbf{x})\omega_2(\mathbf{x}) d\mathbf{x}, \quad \Omega_1, \Omega_2 \in \mathbf{C}_0^m, \tag{18}$$

where $\omega_i(\mathbf{x}) = \dfrac{\partial^m \Omega_i(\mathbf{x})}{\partial x_1 \ldots \partial x_m}$, $i = 1, 2$. Denote by $\mathbf{H}^m$ the closure of the space $\mathbf{C}_0^m$ under the norm $\|\cdot\|$ induced by scalar product (18). For any $m \geq 2$, $\mathbf{H}^m$ is a Hilbert space whose properties are

immediately derived from those for $m = 2$ established in [15]. In particular, the embedding of $\mathbf{H}^m$ into $\mathbf{C}(I^m)$ is compact. Therefore, a function from $\mathbf{H}^m$ equals zero on any "left" side of the cube I^m adjacent to the origin.

Recalling condition **(C2)** imposed on the dependence function Ω_m, consider the problem of minimizing the functional $\int_{I^m} \omega^2(\mathbf{x})d\mathbf{x}$ on the subspace of $\mathbf{H}^m$ specified by the boundary conditions on the "right" sides of I^m adjacent to the point $\mathbf{1} = (1, 1, \ldots, 1)$ provided $\int_{I^m} \Omega(\mathbf{x})d\mu(\mathbf{x}) = 1$, with μ being a finite measure on I^m. In order to describe all possible boundary conditions of this extremal problem we need some notation.

Let $M = \{1, 2, \ldots, m\}$ and let 2^M be the set of all subsets of M. For any $U \subset M$, denote $\mathbf{x}_U$ the $|U|$-dimensional vector $\mathbf{x}_U = (x_i : i \in U)$. Then, any possible set of the boundary conditions has the form

$$\Omega(\mathbf{x})|_{\mathbf{x}_U=\mathbf{1}} = 0, \quad U \in \mathcal{M},$$

where $\mathcal{M} \subset 2^M$ is such that for any $U \subset V \subset 2^M$, $U \in \mathcal{M}$ implies $V \in \mathcal{M}$. That is, if a set U belongs to $\mathcal{M}$, then all its "oversets" also belong to $\mathcal{M}$. The reason for this requirement is simple: if $\Omega \in \mathbf{H}^m$ takes a zero value on the side $\{\mathbf{x}_U = \mathbf{1}\}$, it also takes a zero value on all the subedges of I^m of less dimension.

Remark 1. For any $U \in 2^M$ define an m-dimensional vector of Boolean variables $(y_j = \mathbb{I}(j \in U), j = 1, \ldots, m)$. Then $\mathbb{I}(U \in \mathcal{M})$ is a monotone Boolean function [12]. Denote by $N(m)$ the total number of such functions. Obviously, the number of the above considered extremal problems is also equal to $N(m)$. So far, no explicit formula for $N(m)$ as a function of m has been found. For asymptotic behaviour of $N(m)$ as $m \to \infty$ see [14].

Return to the extremal problem of interest:

$$\|\Omega\|^2_{\mathbf{H}^m} \to \min, \quad \text{where} \quad \int_{I^m} \Omega(\mathbf{x})d\mu(\mathbf{x}) = 1, \tag{19}$$

subject to the conditions

$$\Omega \in \mathbf{H}^m \quad \text{and} \quad \Omega(\mathbf{x})|_{\mathbf{x}_U=1} = 0 \quad \text{for all } U \in \mathcal{M}. \tag{20}$$

For a set $U = (i_1, \ldots, i_l) \in \mathcal{M}$ and its complement (in M) $U^c = (j_1, \ldots, j_k)$, $l + k = m$, put

$$\mathbf{x}_U\mathbf{x}^2_{U^c} = x_{i_1} \ldots x_{i_l}x^2_{j_1} \ldots x^2_{j_k}, \quad \partial\mathbf{x}_U\partial\mathbf{x}^2_{U^c} = \partial x_{i_1} \ldots \partial x_{i_l}\partial x^2_{j_1} \ldots \partial x^2_{j_k},$$

and define the functions

$$K_U(\mathbf{x}, \boldsymbol{\xi}) = K_{i_1}(\mathbf{x}, \boldsymbol{\xi}) \ldots K_{i_l}(\mathbf{x}, \boldsymbol{\xi}), \quad k_{U^c}(\mathbf{x}, \boldsymbol{\xi}) = k_{j_1}(\mathbf{x}, \boldsymbol{\xi}) \ldots k_{j_k}(\mathbf{x}, \boldsymbol{\xi}), \quad \mathbf{x}, \boldsymbol{\xi} \in I^m,$$

where

$$K_j(\mathbf{x}, \boldsymbol{\xi}) = \min(x_j, \xi_j), \quad k_j(\mathbf{x}, \boldsymbol{\xi}) = x_j\xi_j, \quad j = 1, \ldots, m.$$

According to the Lagrange principle applied to a functional on a Banach space (see [1, Sec. 2.2.3]), the necessary condition of a minimum in (19)–(20) is reduced to the Euler–Lagrange equation

$$(-1)^m \lambda \frac{\partial^{2m} \Omega(\mathbf{x})}{\partial x_1^2 \dots \partial x_m^2} = \mu(\mathbf{x}), \tag{21}$$

and the natural boundary conditions

$$\frac{\partial^{l+2k} \Omega(\mathbf{x})}{\partial \mathbf{x}_V \partial \mathbf{x}_{V^c}^2(\mathbf{x})}\Big|_{\mathbf{x}_V=1} = 0, \quad \text{for any } V \notin \mathcal{M},\ V \neq \emptyset, \tag{22}$$

where the Lagrange multiplier l is found from the integral restriction in (19). The following result holds true.

Lemma 3. *Solution to extremal problem* (19)–(20) *is given by the formula*

$$\Omega(\mathbf{x}) = l^{-1} \int_{I^m} \mathcal{G}_{\mathcal{M}}(\mathbf{x}, \boldsymbol{\xi}) d\mu(\boldsymbol{\xi}), \quad \mathbf{x} \in I^m,$$

where $\mathcal{G}_{\mathcal{M}}$ *is the Green function of boundary-value problem* (20)–(22) *equal to*

$$\mathcal{G}_{\mathcal{M}}(\mathbf{x}, \boldsymbol{\xi}) = K_M(\mathbf{x}, \boldsymbol{\xi}) - \sum_{U \in \mathcal{M}} a_U K_{U^c}(\mathbf{x}, \boldsymbol{\xi}) k_U(\mathbf{x}, \boldsymbol{\xi}), \tag{23}$$

with the coefficients a_U *defined recurrently by*

$$\sum_{\substack{V \subset U \\ V \in \mathcal{M}}} a_V = 1, \quad \text{for all } U \in \mathcal{M}, \tag{24}$$

and the constant l *is given by*

$$l = \iint_{I^m \times I^m} \mathcal{G}_{\mathcal{M}}(\mathbf{x}, \boldsymbol{\xi}) d\mu(\mathbf{x}) d\mu(\boldsymbol{\xi}). \tag{25}$$

Proof. First, note that

$$-\frac{\partial^2 K_j(\mathbf{x}, \boldsymbol{\xi})}{\partial x_j^2} = \delta(x_j - \xi_j), \quad \frac{\partial^2 k_j(\mathbf{x}, \boldsymbol{\xi})}{\partial x_j^2} = 0,$$

where δ is the Dirac function. Therefore the function $\mathcal{G}_{\mathcal{M}}$ in (23) satisfies

$$(-1)^m \frac{\partial^{2m} \mathcal{G}_{\mathcal{M}}(\mathbf{x}, \boldsymbol{\xi})}{\partial x_1^2 \dots \partial x_m^2} = \delta(\mathbf{x} - \boldsymbol{\xi}),$$

with an arbitrary choice of the constants a_U. The function $\mathcal{G}_{\mathcal{M}}$ also satisfies natural boundary conditions (22).

Taking into account (20), we arrive at recurrent system (24). Thus, solution to boundary-value problem (20)–(22) is given by (23).

It remains to note that the Lagrange multiplier λ is found from the integral restriction in (19) and has the form (25). The lemma is proved. □

Remark 2. Consider the following three sets of boundary conditions: (i) there are no restrictions on $\Omega \in \mathbf{H}^m$ except for those that specify the space $\mathbf{H}^m$. (ii) $\Omega \in \mathbf{H}^m$ equals zero on any $(m-1)$-dimensional side of I^m, and (iii) $\Omega \in \mathbf{H}^m$ equals zero at the point $\mathbf{1} = (1, \ldots, 1)$. Then $\mathcal{M} = \emptyset$, $\mathcal{M} = 2^M$, and $\mathcal{M} = \{M\}$, respectively, and by Lemma 3 the corresponding Green functions are

$$\begin{aligned} \mathcal{G}_{\emptyset}(\mathbf{x}, \boldsymbol{\xi}) &= \prod_{j=1}^{m} K_j(\mathbf{x}, \boldsymbol{\xi}), \\ \mathcal{G}_{2^M}(\mathbf{x}, \boldsymbol{\xi}) &= \prod_{j=1}^{m} (K_j(\mathbf{x}, \boldsymbol{\xi}) - k_j(\mathbf{x}, \boldsymbol{\xi})), \\ \mathcal{G}_{M}(\mathbf{x}, \boldsymbol{\xi}) &= \prod_{j=1}^{m} K_j(\mathbf{x}, \boldsymbol{\xi}) - \prod_{j=1}^{m} k_j(\mathbf{x}, \boldsymbol{\xi}). \end{aligned}$$

These are covariance functions of the classical Gaussian random fields. They correspond to a Brownian sheet, a Brownian pillow, and a "pinned" Brownian sheet, respectively, that emerge as limiting processes in nonparametric testing of multivariate independence. For example, in the case $m = 2$, the functions $\mathcal{G}_{2^M}$ and $\mathcal{G}_M$ appeared in connection with finding the approximate Bahadur efficiency of independence tests based on the comparison of the multivariate empirical cdf F_n with the product of margins $\prod_{j=1}^{m} F_j$ and with the product of empirical margins $\prod_{j=1}^{m} F_{j,n}$ (see [16, Ch. 5] for details).

4.3 Most favourable alternative to independence

In order to determine the "optimal" distribution function for the sequence $\{S_{m,n}\}_{n\geq 1}$ we have to solve a variational problem of minimization of the functional $\int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x}$ on a set of functions of special type (see inequality (14)). Optimality conditions for the test statistic $S_{m,n}$ are given by the following theorem.

Theorem. *Let $F_\theta \in \mathcal{F}_m$. Then the sequence of test statistics $\{S_{m,n}\}_{n\geq 1}$ is Pitman optimal if and only if*

$$\Omega_m(\mathbf{x}) = C \prod_{j=1}^{m} x_j \left(\prod_{j=1}^{m} (2 - x_j) + \sum_{j=1}^{m} x_j - (m+1) \right), \tag{26}$$
$$\mathbf{x} = (x_1, \ldots, x_m) \in I^m, \quad C > 0.$$

Proof. The test based on $S_{m,n}$ is the "best" for those dependence functions Ω_m that deliver equality in inequality (14). Thus, we minimize the functional $\int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x}$ on the space $\mathbf{H}^m$ subject to

$$\int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x} = 1, \quad \Omega_m(\mathbf{x})|_{x_{i_1} = \ldots = x_{i_{m-1}} = 1} = 0, \quad 1 \leq i_1 < \ldots < i_{m-1} \leq m,$$

where the second constraint on Ω_m is a consequence of condition **(C2)**. Therefore, with the notation of Section 4.2

$$\mathcal{M} = \{M, M \setminus \{1\}, M \setminus \{2\}, \ldots, M \setminus \{m\}\},$$

and the problem (20)–(22) takes the form

$$(-1)^m l \frac{\partial^{2m} \Omega_m(\mathbf{x})}{\partial x_1^2 \ldots \partial x_m^2} = 1,$$

$$\int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x} = 1, \quad \Omega_m(\mathbf{x})|_{x_j=0} = 0, \quad j = 1, \ldots, m,$$

$$\Omega_m(\mathbf{x})|_{x_{i_1}=\ldots=x_{i_{m-1}}=1} = 0, \quad 1 \le i_1 < \ldots < i_{m-1} \le m,$$

$$\frac{\partial^{2m-1} \Omega_m(\mathbf{x})}{\partial x_{i_1} \partial x_{i_2}^2 \ldots \partial x_{i_m}^2} \Big|_{x_{i_1}=1} = 0, \quad 1 \le i_1 \le m,$$

$$\frac{\partial^{2m-2} \Omega_m(\mathbf{x})}{\partial x_{i_1} \partial x_{i_2} \partial x_{i_3}^2 \ldots \partial x_{i_m}^2} \Big|_{x_{i_1}=x_{i_2}=1} = 0, \quad 1 \le i_1 < i_2 \le m,$$

$$\vdots$$

$$\frac{\partial^{m+2} \Omega_m(\mathbf{x})}{\partial x_{i_1} \ldots \partial x_{i_{m-2}} \partial x_{i_{m-1}}^2 \partial x_{i_m}^2} \Big|_{x_{i_1}=\ldots=x_{i_{m-2}}=1} = 0, \quad 1 \le i_1 < \ldots < i_{m-2} \le m.$$

According to Lemma 3 the minimum of $\int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x}$ is attained for the function

$$\Omega_m(\mathbf{x}) = l^{-1} \int_{I^m} \mathcal{G}(\mathbf{x}, \boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{27}$$

where

$$\mathcal{G}(\mathbf{x}, \boldsymbol{\xi}) = \prod_{j=1}^m K_j(\mathbf{x}, \boldsymbol{\xi}) - \sum_{j=1}^m \left(K_j(\mathbf{x}, \boldsymbol{\xi}) \prod_{i \ne j} k_i(\mathbf{x}, \boldsymbol{\xi}) \right) + (m-1) \prod_{j=1}^m k_j(\mathbf{x}, \boldsymbol{\xi}),$$

with K_j and k_j as before. By homogeneity of inequality (14) the extremal function is defined up to a positive constant. Integrating in (27) yields (26). □

Remark 3. For the Spearman-type test statistics $W_{m,n}$ and $V_{m,n}$ the most favourable alternatives are specified by the dependence functions (see [22, Sec. 5])

$$\begin{aligned} \Omega_{m,W}(\mathbf{x}) &= C \prod_{j=1}^m x_j \left(\prod_j x_j - \sum_j x_j + (m-1) \right), \quad \mathbf{x} \in I^m, \quad C > 0, \\ \Omega_{m,V}(\mathbf{x}) &= C \prod_{i=j}^m x_j \sum_{i<j} (1-x_i)(1-x_j), \quad \mathbf{x} \in I^m, \quad C > 0, \end{aligned}$$

respectively. The function $\Omega_{m,V}$, that corresponds to the pair-wise average Spearman's statistic $V_{m,n}$, determines an m-variate extension of the Farlie–Gumbel–Morgenstern distribution introduced in [10, Sec. 5.1].

4.4 Examples

Now we examine, for several choices of cdf F_θ, the Pitman efficiency of $S_{m,n}$ compared to the other two multivariate Spearman-type test statistics, $W_{m,n}$ and $V_{m,n}$.

Example 1. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent copies of the equicorrelated random Gaussian vector $\mathbf{X} = (X_1, \ldots, X_m)$ with $\mathbf{E}X_i = 0$, $\mathbf{Var}X_i = 1$, and $\mathbf{Cov}(X_i, X_j) = \theta$, $1 \le i \ne j \le m$, so that the experiment $\{\mathbf{P}_\theta^n : \theta \ge 0\}$ is normal. The first two terms of Taylor's expansion of the cdf of $\mathbf{X}$ around θ are of the form (7) with the dependence function [5], [22]

$$\Omega_m(\mathbf{x}) = \sum_{1 \le i < j \le m} \varphi(\Phi^{-1}(x_i))\varphi(\Phi^{-1}(x_j)) \prod_{k \ne i,j} x_k, \quad \mathbf{x} \in I^m.$$

The mixed derivative of $\Omega_m(\mathbf{x})$ is

$$\omega_m(\mathbf{x}) = \sum_{1 \le i < j \le m} \Phi^{-1}(x_i)\Phi^{-1}(x_j),$$

and

$$\begin{aligned} \int_{I^m} x_i x_j \, \omega_m(\mathbf{x}) d\mathbf{x} &= 1/(4\pi), \quad \int_{I^m} \omega_m^2(\mathbf{x}) d\mathbf{x} = m(m-1)/2, \\ \int_{I^m} \prod_i x_i \, \omega_m(\mathbf{x}) d\mathbf{x} &= \int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x} = m(m-1)/(2^{m+1}\pi). \end{aligned}$$

Now applying (15)–(17) we obtain

$$e_S(\Omega_m) = e_W(\Omega_m) = \frac{m(m-1)}{2\pi^2\left((4/3)^m - (m/3) - 1\right)}, \quad e_V(\Omega_m) = \frac{9}{\pi^2} \approx 0.9119;$$

The asymptotic efficiency of $S_{m,n}$ and $W_{m,n}$ decreases in m and equals 0.8207, 0.7349, 0.6548 for $m = 3$, 4, 5, respectively, whereas the asymptotic efficiency of $V_{m,n}$ is a constant close to 1 and independent of m. Thus, in the normal case, the average test based on $V_{m,n}$ is asymptotically more efficient than the multivariate Spearman's tests based on $S_{m,n}$ and $W_{m,n}$.

Example 2. Consider the multivariate extension of the Farlie–Gumbel–Morgenstern distribution for which the dependence function is

$$\Omega_m(\mathbf{x}) = \prod_{i=j}^{m} x_j \sum_{i<j} (1 - x_i)(1 - x_j), \quad \mathbf{x} \in I^m.$$

In this case, the average pair-wise Spearman's test based on $V_{m,n}$ is Pitman optimal, i.e., $e_V(\Omega_m) = 1$ (see [22, Sec. 5]). The mixed derivative of $\Omega_m(\mathbf{x})$ is

$$\omega_m(\mathbf{x}) = 1 - \frac{4}{m} \sum_j x_j + \frac{8}{m(m-1)} \sum_{i<j} x_i x_j$$

and

$$\int_{I^m} \omega_m^2(\mathbf{x}) \, d\mathbf{x} = \frac{2}{9m(m-1)}, \quad \int_{I^m} \Omega_m(\mathbf{x}) d\mathbf{x} = \int_{I^m} \prod_{j=1}^{m} x_j \, \omega_m(\mathbf{x}) \, d\mathbf{x} = \frac{1}{9 \cdot 2^m}.$$

Therefore, according to (15) and (16)

$$e_S(\Omega_m) = e_W(\Omega_m) = \frac{m(m-1)}{18((4/3)^m - (m/3) - 1)}. \tag{28}$$

Again, the test statistics $S_{m,n}$ and $W_{m,n}$ are equally efficient in the Pitman sense. Their asymptotic efficiency decreases as m increases, and equals 0.9000, 0.8060, 0.7181 for $m = 3, 4, 5$, respectively.

In both examples the asymptotic equivalence (in the sense of Pitman) of the tests based on $S_{m,n}$ and $W_{m,n}$ is explained by the fact that the corresponding cdfs in model (7) are radially symmetric, i.e., $F_\theta(\mathbf{x}) = \bar{F}_\theta(1-\mathbf{x})$, in which case $s_m(F_\theta)$ and $w_m(F_\theta)$ are known to be equal (see [20, Sec. 3]).

Acknowledgments

The research of A. Nazarov was partly supported by RFBR grant 07-01-00159. The research of N. Stepanova was supported by an NSERC grant. We would like to thank Prof. Yu. V. Tarannikov for communicating us references [12] and [14].

References

[1] V. M. Alexeev, V. M. Tikhomirov, S. V. Fomin, *Optimal Control*. Nauka, Moscow, 1979. (In Russian).

[2] K. Behnen, *Asymptotic optimality and ARE of certain rank order tests under contiguity,* Ann. Mathem. Statist., 42 (1971) 325–329.

[3] A. A. Borovkov, *Mathematical Statistics*, Gordon and Breach Science Publishers, 1998.

[4] D. J. G. Farlie, *The performance of some correlation coefficients for a general bivariate distribution*, Biometrika, 47 (1960) 307–323.

[5] C. Genest, J.-F. Quessy, B. Rémillard, *Asymptotic local efficiency of Cramér–von Mises type tests for multivariate dependence*, Ann. Statist., 35 (2007) 166–191.

[6] G. Gregory, *On efficiency and optimality of quadratic tests,* Ann. Statist., 8 (1980) 116–131.

[7] T. P. Hettmansperger, *Statistical inference based on ranks*, Wiley, New York, 1984.

[8] I. A. Ibragimov, R. Z. Has'minskii, *Statistical Estimation — Asymptotic Theory*, Springer-Verlag, New York, 1981.

[9] H. Joe, *Multivariate Concordance*, J. Multivariate Anal., 35 (1990) 12–30.

[10] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, London, 1997.

[11] M. G. Kendall, *Rank correlation methods*, Griffin, London, 1970.

[12] D. Kleitman, *On Dedekind problem: the number of monotone Boolean functions,* Proc. Amer. Math. Soc., 21 (1969) 677–682.

[13] V. S. Korolyuk, Yu. V. Borovskikh, *Theory of U-statistics,* Kluwer, Dordrecht, 1993.

[14] A. D. Korshunov, *On quantity of monotone functions.* Problems of Cybernetics, Moscow, Nauka, 38 (1981) 5–108. (In Russian.)

[15] A. I. Nazarov, Ya. Yu. Nikitin, *Some extremal problems for Gaussian and empirical random fields.* Amer. Math. Soc. Transl., Ser. 2, ed. by N. N. Uraltseva, 205 (2002) 189–202. (Originally published in Proc. St. Petersburg Math. Soc., 8 (2000) 214–230.)

[16] Ya. Yu. Nikitin, *Asymptotic Efficiency of Nonparametric Tests,* Cambridge University Press, 1995.

[17] Ya. Yu. Nikitin, A. G. Pankrashova, *Bahadur efficiency and local asymptotic optimality of certain nonparametric tests for independence,* J. Soviet Math. 52, No. 2 (1990) 2942–2955. (Originally published in Zap. Nauchn. Sem. LOMI, 166 (1988) 112–127.)

[18] J.-F. Quessy, *Theoretical efficiency comparisons of independence tests based on multivariate vesrions of Spearman's rho*, Metrika. DOI 10.1007/s00184–008–0194–3.

[19] F. H. Ruymgaart, M. C. A. van Zuijlen, *Asymptotic normality of multivariate linear rank statistics in the non-i.i.d. case*, Ann. Statist., 6 (1978) 588–602.

[20] F. Schmid, R. Schmidt, *Multivariate extensions of Spearman's rho and related statistics*, Statist. Probab. Lett., 77 (2007) 407–416.

[21] C. Spearman, *The proof and measurement of association between two things*, Amer. J. Psychol., 15 (1904) 72–101.

[22] N. A. Stepanova, *Multivariate rank statistics for independence and their asymptotic efficiency*, Math. Methods Statist., 12, No. 2 (2003) 197–217.

[23] A. W. van der Vaart, *Asymptotic Statistics*. Cambrige University Press, 1998.